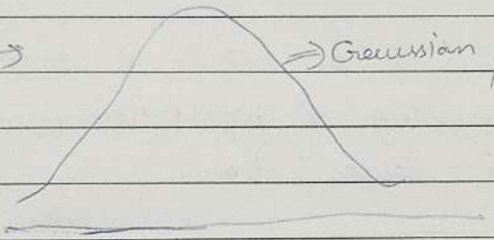
DAY-4 STATS

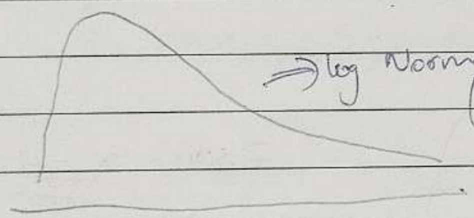
1. Central Limit Theorem
2. Probability
3. Permutation & combination
4. Covariance, Pearson Correlation, Spearman Rank Correlation
5. Bernoulli Distribution
6. Binomial Distribution
7. Power law {Parito Distribution}

CENTRAL LIMIT THEOREM

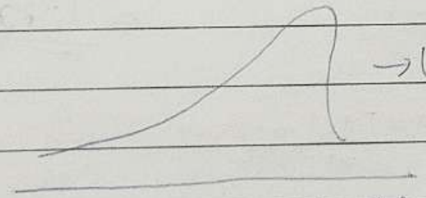
$\boxed{N}$ →



→ Gaussian / Normal Distribution



→ log Normal Distribution (Right skewed)



→ left skewed Distribution

$\boxed{n}$ → size of sample data $\boxed{m}$ → No. of samples

$\boxed{n \geq 30}$

$S_1 \rightarrow \{x_1, x_2, x_3 \;.\;.\;.\;.\; x_n\} \rightarrow \bar{x_1}$ OR $= \bar{S_1}$

$S_2 \rightarrow \{x_3 \cdot x_4 \;.\;.\;.\;.\; x_n\} \rightarrow \bar{x_2}$ $\bar{S_2}$

$S_3 \rightarrow \{x_4, x_1 - - - - - x_n\} \rightarrow \bar{x_3}$

$\vdots$

$\bar{x_m}$

# Central Limit Theorem

for population

→ let $X$ be a random variable with any kind of distribution & we take sample mean with $\boxed{n \geq 30}$ then the plot of all the $m$ sample means would be a Gaussian Distribution.

Wiki - The central limit theorem states that if you have a population with mean $\mu$ & standard deviation $\sigma$ and take sufficiently large random samples from the population with replacement, then the distribution of sample means will be approximately normally distributed.

$\boxed{m}$ → larger the value, better the result.

$m \uparrow\uparrow\uparrow$ more inclined to Gaussian Distribution.

Eg:- Avg size of Shark → If we follow central limit theorem and plot distribution of sample means we can make assumptions

② Probability :- Probability is a measure of the likelihood of an event

Eg:- Tossing a fair coin $P(H) = 0.5$ $P(T) = 0.5$

Rolling a Dice $P(1) = \frac{1}{6}$, $P(2) = \frac{1}{6}$, $P(3) = \frac{1}{6}$

i) Mutual Exclusive event :- 2 events are mutually exclusive, if they cannot occur at the same time.
Eg:- Tossing a coin, Rolling a dice

ii) Non Mutual Exclusive Events :- Two events can occur at the same time. Eg:- Picking randomly a card from a deck of card. Two events "heart" and "king" can be selected.

Mutual Exclusive Event

Q. What is the probability of coin landing on heads or tails.

$\Downarrow$

Addition rule for mutual exclusive events.

$$P(A \text{ or } B) = P(A) + P(B)$$
$$= \frac{1}{2} + \frac{1}{2} = 1$$

Q. What is the probability of getting 1 or 6 or 3 while rolling a dice?

$$P(1 \text{ or } 6 \text{ or } 3) = P(1) + P(6) + P(3) = 3 \times \frac{1}{6} = 0.5$$

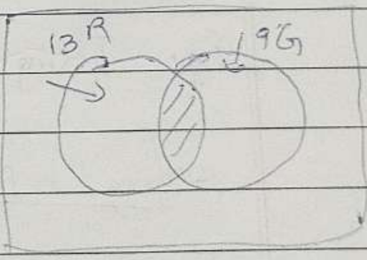
Non-mutual Exclusive Events

Q. Bag of Marbles : 10 Red, 6 Green, 3 (R & G)
When picking randomly from a bag of marbles what is the probability of choosing a marble that is red & Green?

$\Downarrow$

Addition rule for non-mutual exclusive event

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
$$= \cancel{\frac{10}{19}(R) + \frac{6}{19}(G)} - \frac{3}{19}$$




$$= \frac{13}{19}(R) + \frac{9}{19}(G) - \frac{3}{19}$$

$$= \frac{19}{19} = 1$$

Q. Deck of cards → What is the probability of choosing ♡ Or Queen at the same time ?

Ans $P(H) + P(Q\spadesuit) - P(H \& Q) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$

$= \frac{4}{13}$

* Multiplication Rule

Dependent Rules : Two events are dependent if they affect one another.

Bag of marbles $\left\{ \begin{array}{l} \boxed{O\ O\ O\ O} \rightarrow W \\ \boxed{O\ O\ O} \rightarrow Y \end{array} \right.$

$\Rightarrow P(W) = \frac{4}{7} \longrightarrow P(Y) = \frac{3}{6}$

↑

white marbles ⟶ 1

Q What is the probability of rolling and "5" and then a "3" with the a normal 6 sided dice. {Independent event

Ans $P(1) = \frac{1}{6}$  $P(2) = \frac{1}{6}$ . $P(3) = \frac{1}{6}$  $P(4) = \frac{1}{6}$

$P(A \text{ and } B) = P(A) \times P(B)$

$= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

$$P(A \text{ or } B) \Rightarrow \begin{cases} \rightarrow \text{Mutual Exclusive} \\ \rightarrow \text{Non Mutual Exclusive} \end{cases}$$

$P(A \text{ or } B) = P(A) + P(B) - \boxed{P(A \text{ and } B)}$ → Non mutual exclusive

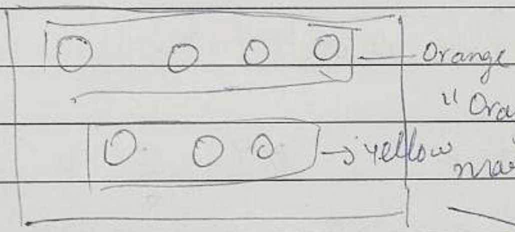
$P(A \text{ or } B) = P(A) + P(B)$ → Mutual Exclusive

Dependent & Independent Events

Tossing a coin

$P(H) \times P(T) = 0.25$

$P(A \text{ and } B) = P(A) \times P(B)$

② 


Probability of drawing a "Orange" and then drawing a "yellow" marble from the bag?

→ Dependent event.

Ans. $P(O) = \frac{4}{7}$ , $P(Y) = \frac{3}{6}$

$P(Y/O) = \frac{3}{6}$ → Probability of Yellow given Orange has happened

→ $\boxed{\text{Conditional Probability}}$

$P(O \text{ and } Y) = P(O) \times \boxed{P(Y/O)}$

$= \frac{4}{7} \times \frac{3}{6} = \frac{4}{14} = \frac{2}{7}$ //

$$\boxed{Cov(X, Y) = \frac{\Sigma (x_i - \bar{x}) \times (y_i - \bar{y})}{n-1}}$$

$\Downarrow$

$\simeq$

$$\sigma^2 = \frac{\Sigma (x_i - \bar{x})}{n-1}$$

$$Cov(X, X) \Leftarrow \boxed{s^2 = \frac{\Sigma (x_i - \bar{x}) \times (x_i - \bar{x})}{n-1}}$$

$$\boxed{Cov(X, X) = Var\ X}$$

| X Age | ~~Y~~ X wt. | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x}) \times (y_i - \bar{y})$ |
|---|---|---|---|---|
| 12 | 40 | -3 | -11 | 33 |
| 13 | 45 | -2 | -6 | 12 |
| 15 | 48 | 0 | -3 | 0 |
| 17 | 60 | 2 | 9 | 18 |
| 18 | 62 | 3 | 11 | 33 |
| | | | | 66 + 30 = 96 |

$$Cov. = \frac{96}{5-1} = 24$$

$\downarrow$

Whenever there is +ve Covariance

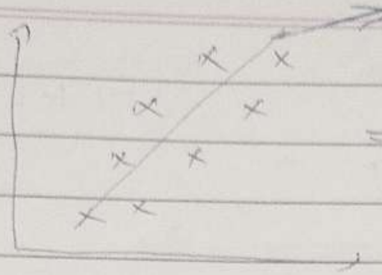
| | |
|---|---|
| X ↑ | Y ↑ |
| X ↓ | Y ↓ |

If negative, -ve Covariance

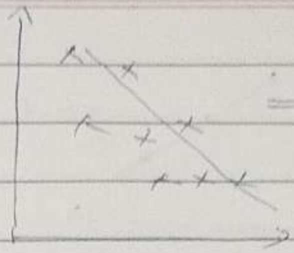
| | |
|---|---|
| X ↑ | Y ↓ |
| X ↓ | Y ↑ |

Covariance = 0 → No relationship b/w X & Y

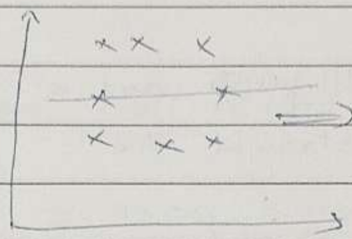
This called best fit line



⟹ +ve



⟹ -ve Covariance



⟹ No relation

Eg:-

| x | y |
|---|---|
| 10 | 5 |
| 8 | 14 |
| 7 | 30 |
| 6 | 33 |

$Cov(x, y) = \underline{-ve}$

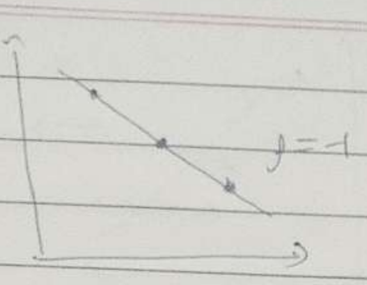
* Pearson Correlation Coefficient

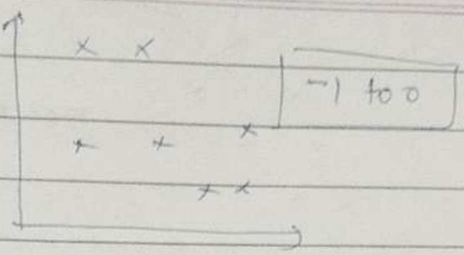
$$\boxed{\rho(x, y) = \frac{Cov(x, y)}{\sigma_x * \sigma_y}}$$

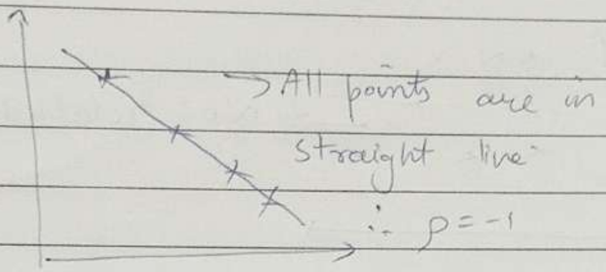
Why this is needed → In covariance there is no such limit as to how much positive or -ve value can be.

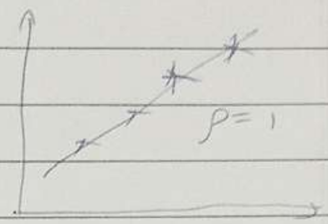
In case of Pearson correlation coefficient it is [-1 — 1]

More the value towards +1 more positive correlated
More " " -1 more negatively "

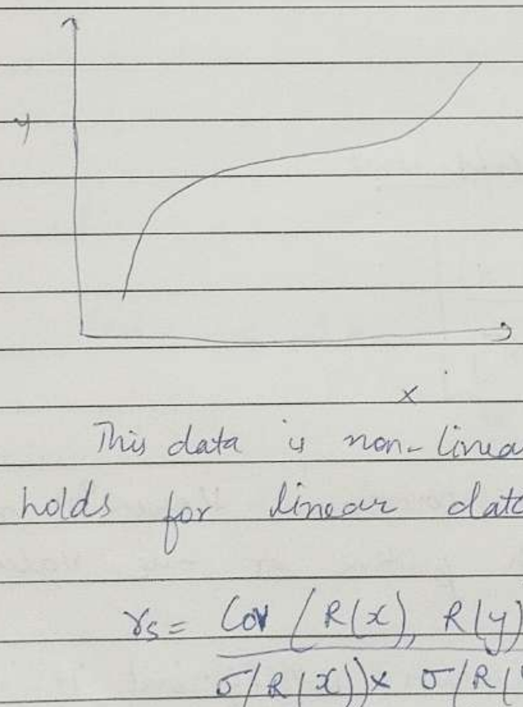



# Spearman Rank Correlation




Spearman correlation = 1
Pearson correlation = 0.88

This data is non-linear, Pearson correlation only holds for linear data.

$$r_s = \frac{\text{Cov}\left(R(x), R(y)\right)}{\sigma\left(R(x)\right) \times \sigma\left(R(y)\right)}$$

$R(x) \rightarrow$ Rank of $x$
$R(y) \rightarrow$ Rank of $y$

In ascending order

| X | Y | R(X) | R(Y) |
|---|---|---|---|
| 10 | 4 | 4 | 1 |
| 8 | 6 | 3 | 2 |
| 7 | 8 | 2 | 3 |
| 6 (Smallest) | 10 | 1 | 4 |

Why this correlation will be used?

X    Y    Z    O/P)

* If X & O/P are highly +vely or -vely correlated it is good.

* If Y = 0.2 with O/P then we can drop this value.

* If X & Y are correlated with 0.95 we can drop one of the column.

Input feature: Exp  Degree  City

O/P feature: Salary

→ We can drop degree since exp & degree are not correlated. If in some ways they are related we will not drop it.

→ If Exp & Degree 95% we can drop one column.